Minseo Kim
DS210
Final Project
December 15, 2023

Project Overview:

The “Wine Quality” dataset, in red or white wine versions, contains information about physiochemical properties and sensory characteristics of Portuguese Vinho Verde wines. My project focuses on analyzing the data, involving data preprocessing, graph creation, dataset splitting, and classification using K-Nearest Neighbors (KNN) and Decision Tree algorithms.

1. **preprocessing.rs**

- Load wine quality data from a `winequality.csv` file
- Clean it by filling null values
- Transform categorical value to numeric
    - In the “Wine Quality” dataset, the only non-numerical value was “color” which was represented by “red” or “white”. Here, I transformed the categorical values to “1” and “2” each.

2. **graph.rs**

- Generate graphs depicting the relationship between wine quality and alcohol content for both red and white wines using fetched data from UCI's URLs.
    - Since I compiled a csv file of red wines and white wines for classification analysis that were separated at first, I tried fetching data from the website instead of using the data I downloaded and compiled.
    - I chose to depict the relationship between wine quality and alcohol content since I was curious if stronger wine(wine with higher percentage of alcohol) is evaluated as a good quality of wine.

3. **split_dataset.rs**

- Split the cleaned dataset into training and testing sets for model training and evaluation.

4. **classification.rs**

- Implement KNN and Decision Tree classifiers from smartcore to predict wine quality based on provided features.

Output:

The output of this project includes:

- **Graphs**: `red_wine_quality.png` and `white_wine_quality.png` depicting the relationship between wine quality and alcohol content for red and white wines, respectively.
- **Cleaned Data**: The cleaned dataset after preprocessing, displaying there's no missing values in the dataset.
- **Classification Accuracy**: The accuracy scores of the KNN and Decision Tree models for predicting wine quality based on the provided dataset.
    - **K-Nearest Neighbors (KNN)**: Achieved an accuracy of approximately 48.34%.
    - **Decision Tree**: Achieved an accuracy score of around 58.66%.
- **Testing:** The code includes tests for data loading and graph creation modules (`test_data_loading` and `test_graph_creation` functions). Tests were performed successfully, which means the data and graph have successfully loaded and created, respectively.

Analysis:

1. **K-Nearest Neighbors (KNN):**

The KNN model achieved an accuracy score of approximately 48.34%, indicating that the KNN model correctly predicted wine quality labels for nearly half of the test dataset. Depending on the large number of classes and data distribution, this accuracy might be reasonable. It also suggests that the relationships between features and wine quality might not be linear or easily discernible using simple distance-based classification. I believe that experimentation with different distance metrics, optimal 'k' values, or feature scaling could potentially enhance KNN's performance.

2. **Decision Tree:**

The Decision Tree model achieved an accuracy score of around 58.66%. It outperforms KNN classification accuracy, indicating that iIt correctly predicted wine quality labels for more than half of the test dataset. I believe this is because Decision Trees can capture nonlinear relationships between features and target variables. However, the Decision Tree model have

limitations in capturing complex interactions between features that could influence wine quality.

In the future …

I would use cross-validation techniques and conduct hyperparameter tuning to validate the stability and robustness of classification models. Especially for the Decision Tree model, I would analyze feature importance from the model to understand which features contribute most significantly to predicting wine quality. For furthur analysis, I could explore ensemble methods like Random Forests or Gradient Boosting, which often improve predictive performance by combining multiple models. Not only using this “Wine Quality” data, I can explore additional data and analyze in-depth or perform feature engineering to create more informative features.